Published in: Dipper, Stefanie/Neubarth, Friedrich/Zinsmeister, Heike (eds.): Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016). Bochum: Ruhr-Universität Bochum, 2016. Pp. 38-43.
(Bochumer Linguistische Arbeitsberichte 16)

Creating an extensible, levelled study corpus of Russian

Dolores Batinić
URPP Language and Space
University of Zurich
dolores.batinic@uzh.ch

Sandra Birzer
Institute of Slavonic Studies
University of Innsbruck
sandra.birzer@uibk.ac.at

Heike Zinsmeister
Institute of German Studies
University of Hamburg
heike.zinsmeister@uni-hamburg.de

Abstract

In this paper, we present first results of training a classifier for discriminating Russian texts into different levels of difficulty. For the classification we considered both surface-oriented features adopted from readability assessments and more linguistically informed, positional features to classify texts into two levels of difficulty. This text classification is the main focus of our Levelled Study Corpus of Russian (LeStCoR), in which we aim to build a corpus adapted for language learning purposes – selecting simpler texts for beginner second language learners and more complex texts for advanced learners. The most discriminative feature in our pilot study was a lexical feature that approximates accessibility of the vocabulary by the second language learner in terms of the proportion of *familiar* words in the texts. The best feature setting achieved an accuracy of 0.91 on a pilot corpus of 209 texts.

1 Introduction

Selecting texts of an appropriate difficulty level is a challenging task for both teachers of a second language (L2) as well as the learners themselves. This becomes particularly evident when learners are working with linguistic corpora which is part of many foreign language studies in the digital age (Römer, 2008; Steinbach and Birzer, 2011): Linguistic corpora do not normally differentiate between texts suitable for beginner and more advanced L2 learners.

One way to deal with text selection for L2 learning purposes is simplifying texts (Karpov and Sibirtseva, 2014; Vajjala and Meurers, 2014), another one is compiling texts selected for different proficiency levels as an additional resource for learners especially on a beginner and intermediate level (Cobb, 2007; Allan, 2009). This paper contributes to the second line of research. In this paper, we introduce our concept for creating a Levelled Study Corpus of Russian (LeStCoR) stratified into texts suitable for L2 learners of different proficiency levels. While the sampling and creation of LeStCoR is still work in progress, we will mainly focus on one aspect: the method of automatically classifying Russian texts according to the difficulty they pose for L2 learners. Since our goal is to provide an extensible study corpus of Russian, we need a tool that supports the classification of new texts in an efficient and consistent way. To this end, we train a classifier on manually labelled texts and use surface as well as linguistically motivated features to discriminate between simple (Class I) and more difficult texts (Class II).

It is important to note that in our approach automatic classification is used by the corpus creator – not the learners themselves – to identify texts with an appropriate difficulty level for integrating them into the corpus. The classification is seen as a preprocessing step followed by additional manual checking if deemed necessary. This means that the classification is performed 'behind the scenes' in terms of Aston (2000). It is not offered 'on stage' as a method for learners to identify appropriate texts by themselves (Vajjala and Meurers, 2013).

The paper is structured as follows. In Section 2, we introduce related work on classifying texts automatically according to their difficulty. Section 3 describes the target text selection. In Section 4, we introduce characteristics that are indicative for text difficulty and detail how we operationalized them as features. Section 5 describes the actual feature selection. In Section 6, we evaluate our approach by a pilot study performed on 209 texts that demonstrates the applicability of the classification method. We close with a discussion of the results and further work.

2 Related Work

There is a long tradition of assessing the difficulty of a text in terms of surface-oriented readability measures that allow the researchers to compare different texts in an objective way (see Dubay (2004) for a historical overview). In addition to the classical surface-oriented measures that mainly take simple word counts, word and sentence lengths etc. into account, other approaches integrate lexical, syntactic, and discourse features that address the lexical coverage of a text, its parts of speech, syntactic structures, and cross-sentential features like the referential overlap and relations between clauses triggered by discourse connectives (McNamara et al., 2014; Napolitano et al., 2015). Machine learning approaches make use of the fact that different measures quantify different aspects of the text difficulty characteristics (Schwarm and Ostendorf, 2005; Vajjala and Meurers, 2012; Karpov and Sibirtseva, 2014). Many related works focused on establishing the level of text difficulty for native speakers (Collins-Thompson and Callan, 2004; Schwarm and Ostendorf, 2005; Feng et al., 2010). However, studies of the difficulty level for L2 learners have also been conducted recently, with the underlying hypothesis that text comprehensibility is perceived very differently by L2 learners (François, 2014; Heilman et al., 2007; Xia et al., 2016).

3 Compilation of a seed corpus

LeStCoR is intended to grow over time by being extended with new texts. For the pilot study on text classification, we selected 209 texts from the Test of Russian as a Foreign Language (TORFL, Russian: TRKI) reading and listening tasks. The pilot corpus is stratified into two classes: Class I contains 136 texts that belong to beginners' or lower intermediate levels (TRKI levels elementary, basis and level 1), whereas Class II contains the other 73 texts of intermediate or advanced levels (TRKI levels 2, 3 and 4). Table 1 gives an overview of the text distribution across the TRKI levels and our text difficulty classes (I & II). We also provide the corresponding levels of the Common European Framework of References for Languages (CEFR) for comparison.

As shown in Table 1, the distribution of texts per class was not homogeneous, since we were able to provide more texts for Class I than for Class II. Some of the texts needed to be OCRed and manually corrected. All texts were part-of-speech tagged and lemmatized with TreeTagger (Schmid, 1994) using parameter files trained on the disambiguated version of the Russian National Corpus (Plungian, 2005; Plungian et al., 2009; Sharoff et al., 2008).

Class	TRKI	CEFR	Sem	#Texts
I	elementary	A1	1st	43
	basis	A2	2nd	43
	1	B1	2nd	50
II	2	B2	3rd	38
	3	C1	4th	30
	4	C2	indep	5

Table 1: TRKI proficiency levels and sampling of the pilot corpus (*#Text*: number of texts, *Class*: simple vs. difficult texts; *Sem*: Semester, *indep*: semester independent).

4 Candidates for features

In this pilot study, we mainly focused on surface features that are employed in traditional readability scores and linguistically motivated token-related lexical and morphosyntactic features. For the linguistic features we tested to what extent the proportion of 'familiar' words, the proportion of 'abstract' words and the proportion of different parts of speech in text may be indicative of the text difficulty.

Average readability score. For calculating the readability scores, we adapted the Python implementation of existing readability measures by Rik Goldman[1] to Russian and calculated an average grade score based on seven common measures (for an overview of most scores see DuBay (2004); the Coleman Liau Index Score is described in Coleman and Liau (1975)):[2]

- Flesch-Kincaid Grade Level
- Coleman Liau Index Score
- (Gunning) Fog
- SMOG Index
- Automated Readability Index
- New Dale Chall Adjusted Grade Level[3]
- Powers-Sumner-Kearl Grade Level

[1] Goldman's implementation: `https://github.com/ghoulmann/py-readability-statistics`.

[2] A demo-version of our text difficulty calculator can be accessed at `http://www.lestcor.com/`.

[3] Calculating the New Dale Chall Adjusted Grade Level makes use of the concept of *hard words*. For English this is done by counting words in text not belonging to the Dale Chall list of 3,000 frequent English words. In our adaptation to Russian, we defined 'hard words' in Russian texts as those having four or more syllables.

Readability scores can be interpreted as an estimation of the number of years of education a person has had. An average readability score of 5 indicates that the given text should be easily comprehensible for a fifth-grade student, whereas a score higher than 15 means that the text is best suited for college graduates.

Familiar words. This feature operationalizes the accessibility of the vocabulary by L2 learners. It measures how much of the text is covered by core vocabulary and other words that are easy to grasp by an adult learner. As *core vocabulary* we used the list of 5,000 most frequent Russian lemmas compiled by Sharoff (2002). A core vocabulary of 5,000 most frequent words is expected to enable the learner to understand about 80% of a text (Hiebert and Kamil, 2005). In addition, as familiar words we also considered numerals, proper names, pronouns, and internationalisms. The latter are treated as familiar words because adult learners of Russian may easily understand them without being familiar with the Russian vocabulary itself. Some examples are бокс 'box', бейсбол 'baseball', and телефон 'telephone'. The list of internationalisms was gathered from Wikipedia's list of internationalisms in the Russian language. We assumed that a high proportion of familiar words was indicative for texts with low difficulty.

Abstract words. We calculated the average occurrence of abstract words in sentences by counting the words in a text having typical abstract word endings, such as -изм '-ism', -ость '-ness', -ство '-ship', -ота '-ness', -ание / -ение (markers of nominalized verbs) and dividing it by the total number of sentences in a text. We also experimented with the proportion of abstract words in the whole text. We assumed that abstract words occurred more frequently in sentences from higher classes. We did not discriminate between internationalisms and abstract words so that there is a certain overlap and potential correlation.

Parts of speech. In order to verify if there is a prevalence of a particular part of speech in sentences of Class I and Class II, we considered the average occurrences of nouns, verbs, pronouns, adjectives, adverbs, adpositions, conjunctions, and particles. Relying on the study conducted by Feng et al. (2010), we expected nouns to have a higher predictive power than other parts of speech.

Syntactic and discourse features. With the idea that they could be discriminative for difficult texts, we studied the distribution of adverbial participles, perfect participles, and marking of conditional (чтобы 'in order to').

Content words. We calculated the proportion of nouns, adjectives, verbs, and adverbs in texts. We assumed that a high proportion of content words may be a good indicator of text difficulty: We expected that the more content words per text, the more difficult the text.

Type/token ratio. We calculated the ratio of unique words in texts (types) to the total number of word occurrences (tokens) in texts. A low ratio would indicate a more difficult text due to a high number of different words.

5 Feature selection

Before selecting the actual feature combinations for the classifier, we observed the differences in their distributions within texts of Class I and Class II. As shown in Figure 1, the average proportion of familiar words in texts of Class I differed from the one in Class II (an average text in Class I contained 94% of familiar words, whereas an average text from Class II contained 83% of familiar words). A difference in the two classes was also considerable for the features average readability (per text). Figure 2 shows that the average absolute frequencies of abstract words and nouns per sentence were also discriminative, followed by adjectives and adpositions. In order to find thresholds which would discriminate between Class I and Class II, we first calculated the average distribution of a given feature for each class. Then we experimented with the classification model by setting initially the two averages as thresholds and incrementing/reducing them until we reached the highest accuracy for the given model. We also investigated different readability measures and found that Flesch-Kincaid Score seemed to discriminate between the two groups more strongly than other readability measures, so we used it as a separate feature as well. The proportion of content words and type/token ratio did not prove to be discriminative for Class I and Class II. The same applies to our syntactic and discourse features, which were too infrequent in the selected TRKI texts to play a role in the classification process (for instance, the conditional marker чтобы 'in order to' occurs only four times in Class I and five times in Class II).

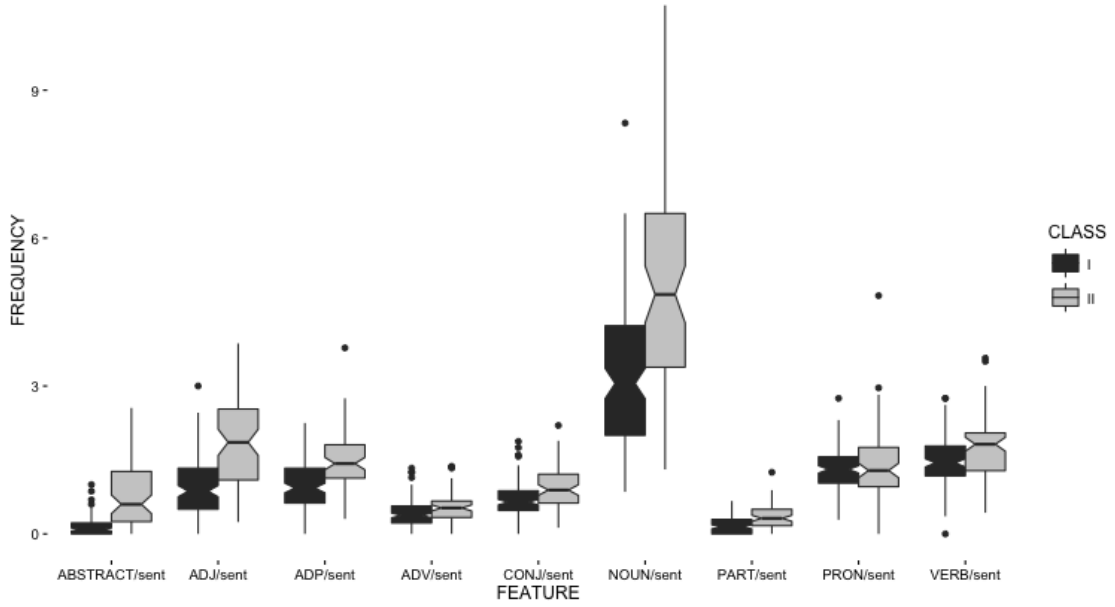

Figure 2: Boxplots for the absolute frequencies of abstract words and different parts of speech (per sentence) in Class I & II. Notches indicate medians and their 95% confidence intervals; dots mark outliers. (Created with R's ggplot2 package).

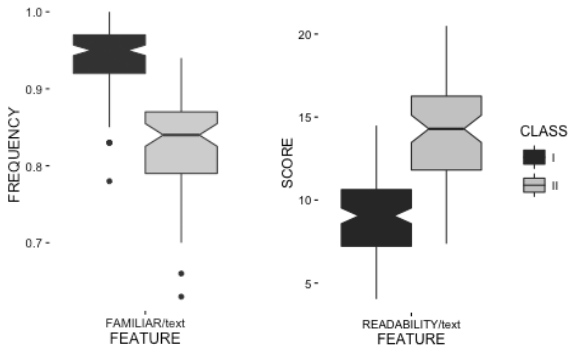

Figure 1: Boxplots for the relative frequencies of familiar words (to the left) and readability scores (to the right) per text in Class I & II.

6 Results and Discussion

We performed a classification with Naive Bayes (NLTK[4], Bird et al., (2009)) and 10-fold cross validation. As a baseline we assumed that the classifier would (randomly always) assign Class I which would result in 65% of the texts being correctly classified on average (136/209). The classifier achieved an accuracy of 0.91 by predicting the text difficulty level by combining the features average readability, familiar words, abstract words, nouns and adjectives. Contrary to our expectations, the average readability score alone did not prove to be discriminative enough (accuracy of 0.64). However, the models that combined average readability with other features reached an accuracy between 0.89 and 0.91 (see M3-M7 in Table 2). Familiar words were highly informative even as a separate feature: When setting the threshold of $> 90\%$ of familiar words per text, the model reached the accuracy of 0.84. This finding suggests that building a two-levelled corpus may be done in a relatively accurate way by using a simple feature such as the proportion of familiar words as basis and extending it with readability scores and more linguistically motivated features.

A low predictability power of the feature average readability score can be related to several factors. Firstly, the average of seven different readability measures smooths the difference between classes which is observed when dealing with particular readability measures separately. For instance, the average Powers-Sumner-Kearl Grade Level for Class II is 9.9, whereas the average Flesch-Kincaid Score for Class II amounts to 18.4. Secondly, different readability measures serve different purposes; for instance, Powers-Sumner-Kearl Grade Level is generally used for children under 10 years. Lastly, for lack of resources we only had five texts repre-

[4]NLTK: `http://www.nltk.org/`.

		Models						
Feature	Threshold	M1	M2	M3	M4	M5	M6	M7
Flesch-Kincaid score	> 19						x	
	< 9							
Average readability	> 15			x	x			
	> 12	x				x	x	x
#Familiar words	$< 80\%$ / t			x	x	x	x	x
	$> 90\%$ / t		x	x	x	x	x	x
#Abstract words	$> 8\%$ / s				x	x	x	x
	$< 2\%$ / s						x	
#Nouns	$> 60\%$ / s					x	x	x
	$< 20\%$ / s					x	x	x
#Adjectives	$> 16\%$ / s					x	x	x
	$< 5\%$ / s					x	x	x
#Adpositions	$> 20\%$ / s					x		
Mean accuracy		.64	.84	.89	.89	.89	.90	.91
sd		$\pm$ .10	$\pm$.08	$\pm$.05	$\pm$.07	$\pm$.05	$\pm$.06	$\pm$.06

Table 2: Classification results with different feature selections. According to a two sample t-test, the accuracies of M2-M7 are significantly different from the ones of M1; M7 differs from M2 with an error probability of $p = 0.05864$.

senting the level TRKI 4. Other texts of this level would presumably have had high average readability scores, which would in consequence ameliorate the prediction strength of this variable.

The proportion of *familiar* words, though, proved to be a well-suited predictor for discriminating between simple and difficult texts for L2 learners. This is likely due to the fact that *familiar* words included not only frequent words, but also numbers, pronouns, internationalisms and named entities, which, although they might still be incomprehensible or difficult to read for L2 learners, they do not compromise their comprehension of the text as a whole. Moreover, a list of the 5,000 most frequent Russian lemmas proved to be a suitable amount of words to use as a threshold for discriminating between texts below and above CEFR's B2 level, corresponding to TRKI 2.

In further work, we plan to work with the core vocabulary for all TRKI levels separately, instead of using the top word frequency list of 5,000 lemmas as a threshold between simple and difficult vocabulary. Once we provide some more text material, we are also planning to include more linguistically motivated features, such as discourse markers and syntactic markers as well as semantic features, such as the proportion of academic vocabulary words (Vajjala and Meurers, 2012; Vajjala and Meurers, 2014). Moreover, we are considering using a language-modelling approach (Collins-Thompson and Callan, 2004), which may be well suited for an extensible corpus.

7 Conclusion

We performed a text classification study to classify original, non-adapted Russian texts into two levels of difficulty for L2 learners. The trained classifier beat the baseline and achieved an average accuracy of 0.91 with surface-oriented features complemented by vocabulary-based features including part of speech information. The list of most frequent Russian words extended with named entities, numbers, pronouns and internationalisms proved to be the best suited predictor for text difficulty classification aimed to L2 learners. More linguistically-motivated features like syntactic and discourse features did not improve the classification results but we expect more conclusive results on a larger training base.

Acknowledgments

We would like to thank the anonymous reviewers for their very helpful remarks and Piklu Gupta for improving our English. All remaining errors are ours.

References

Rachel Allan. 2009. Can a graded reader corpus provide 'authentic' input? *ELT journal*, 63(1):23–32.

Guy Aston. 2000. Learning English with the British National Corpus. In M.P. Battaner and C. L'opez, editors, *VI jornada de corpus lingüístics*, pages 15–40, Barcelona.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*. O'Reilly Media, Inc.

Tom Cobb. 2007. Computing the vocabulary demands of L2 reading. *Language Learning & Technology*, 11(3):38–63.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.

Kevyn Collins-Thompson and James P Callan. 2004. A language modeling approach to predicting reading difficulty. In *HLT-NAACL 2004: Main Proceedings*, pages 193–200, Boston, MA.

William H DuBay. 2004. The Principles of Readability. *Online Submission*: `http://files.eric.ed.gov/fulltext/ED490073.pdf`.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *Proceedings of Coling 2010: Posters*, pages 276–284, Beijing, China.

Thomas François. 2014. An analysis of a French as a foreign language corpus for readability assessment. In *Proceedings of the 3rd workshop on NLP for computer-assisted language learning at SLTC 2014*, Uppsala University.

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proceedings of HLT-NAACL 2007*, pages 460–467, Rochester, New York.

Elfrieda H. Hiebert and Michael L. Kamil. 2005. *Teaching and Learning Vocabulary: Bringing Research to Practice*. Lawrence Erlbaum Associates, Mahwah, New Jersey.

Nikolai Karpov and Vera Sibirtseva. 2014. Towards automatic text adaptation in Russian. *Higher School of Economics Research Paper No. WP BRP*, 16.

Danielle S McNamara, Arthur C Graesser, Philip M McCarthy, and Zhiqiang Cai. 2014. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

Diane Napolitano, Kathleen Sheehan, and Robert Mundkowsky. 2015. Online readability and text complexity analysis with TextEvaluator. In *Proceedings of NAACL 2015: Demonstrations*, pages 96–100, Denver, Colorado.

Vladimir A. Plungian, Ekaterina V. Rakhilina, and Tatjana I. Reznikova, editors. 2009. *Nacional'nyj korpus russkogo jazyka: 2006-2008. Novye rezul'taty i perspektivy*. Nestor-Istorja, St. Petersburg.

Vladimir A. Plungian, editor. 2005. *Nacional'nyj korpus russkogo jazyka: 2003-2005. Rezul'taty i perspektivy*. Indrik, Moscow.

Ute Römer. 2008. Corpora and language teaching. In Anke Lüdeling and Merja Kytö, editors, *Corpus linguistics: An International Handbook*, Handbücher zur Sprache und Kommunikationswissenschaft. Volume 1, pages 112–130. Mouton de Gruyter, Berlin.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Sarah Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of ACL 2005*, pages 523–530, Ann Arbor, Michigan.

Serge Sharoff, Mikhail Kopotev, Tomaz Erjavec, Anna Feldman, and Dagmar Divjak. 2008. Designing and evaluating Russian tagsets. In *Proceedings of LREC 2008*, pages 279–285, Marrakech, Marrocco.

Serge Sharoff. 2002. Meaning as use: Exploitation of aligned corpora for the contrastive study of lexical semantics. In *Proceedings of LREC 2002*, Las Palmas, Spain.

Andrea Steinbach and Sandra Birzer. 2011. Authentisches Sprachmaterial schnell gefunden. Das Potenzial russischer Textkorpora im Russischunterricht. *Praxis Fremdsprachenunterricht*, (2):7–10.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173, Montréal, Canada.

Sowmya Vajjala and Detmar Meurers. 2013. On the applicability of readability models to web texts. In *Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 59–68.

Sowmya Vajjala and Detmar Meurers. 2014. Assessing the relative reading level of sentence pairs for text simplification. In *Procedings of EACL 2014*, pages 288–297, Gothenburg, Sweden.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA.